DOCUMENT RESUME

ED 101 015 TM 004 255

AUTHOR Aiken, Lewis R.; Williams, E. Newsom
TITLE The Effects of Instruction, Option Keying, and Knowledge of Test Material on Seven Methods of Scoring Double-Choice Items (Alternative Scoring Methods for Ordered and Double-Choice Items).
PUB DATE [Aug 74]
NOTE 22p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, D.C., August, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Computer Programs; *Guessing (Tests); *Multiple Choice Tests; *Response Style (Tests); *Scoring Formulas; Statistical Analysis; *Testing; Tests of Significance

ABSTRACT

A generalized approach to scoring ordered multicategory items having one or more options in each category is described first, and differential weighting schemes for scoring dichotomous items are considered. Seven scoring formulas, five of which take into account the fact that the number of options keyed as correct may be different for the two response categories in dichotomous items are presented. The results of an experiment involving the seven methods of scoring a 100-item general psychology test as dependent variables are then given. The four independent variables were (P) proportion of total number of items on which option "a" was keyed as correct, (T) whether examinees were told the proportion of "a" options keyed as correct, (G) whether examinees were instructed to guess when in doubt, and (K) two levels of examinee's knowledge of subject-matter of the test (low vs. high). Results of the statistical analyses are presented. (Author/RC)

TM 004 255 — ED101015

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Effects of Instructions, Option Keying, and Knowledge of Test Material on Seven Methods of Scoring Double-Choice Items (Alternative Scoring Methods for Ordered and Double-Choice Items)

Lewis R. Aiken & E. Newsom Williams

Guilford College

Many investigations of the effects of correction for guessing formulas and differential weighting of item responses on objective test items have been conducted (see Stanley & Wang, 1970). A number of inventive schemes for differential weighting have been studied, and, as with the "correction for guessing" formulas, such weighting procedures can increase the variance of total test scores. But the order of the scores does not change appreciably, and the general finding has been that assigning different numerical weights to each of a group of intercorrelated items has little effect on the reliability and validity of a test. This is especially true when the number of items is large, the criterion factorially simple, and 90% or more examinees complete the test. The implication of this finding is that, at least on long multiple-choice tests, it is usually not worthwhile to "correct for guessing" or assign different scoring weights to different responses to an item. However, the same conclusion does not necessarily hold for true-false, rearrangement, and ordered multicategory items.

Scoring Rearrangement Items

Elsewhere (Aiken, 1970) the senior author has recommended that rather than scoring according to the number of correspondences, ranking or rearrangement items on which the examinee is limited to one response per category be scored by the following formula:

$$(1)\quad S = N(1 - 2D/(N^2 - i)),$$

where S = the examinee's final score on the item, N = the number of categories, and D(Diff) = the sum of the absolute values of the differences between the numbers of the categories in which the examinee places the options and the numbers of the categories in which they correctly belong. When using this formula, the minimum score is 0 and the maximum score is equal to the number of things to be rearranged.

The nomograph in Figure 1 was constructed to facilitate the use of formula 1. A straight line extended from the computed value of Diff on the Diff axis through the appropriate point on the Number axis corresponding to the number of things to be rearranged will intersect the Score axis at the desired score point. Empirical studies of items with 9 and 16 categories (N = 9 & 16, ten samples of size 40) reveal that the correlation between scores obtained from formula 1 and scores obtained simply by counting the number of correspondences between the examinee's arrangement and the correct arrangement are in the high .80s or low .90s. The mean scores obtained from formula 1 are significantly higher than those obtained by counting correspondences, but the standard deviations are slightly higher in the latter case.

Generalizing from the conventional rearrangement item with its restriction to one response per category, an appropriate scoring formula for ordered multicategory items having no limit on the number of items to be placed in any category is:

$$(2)\quad S = N(1 - D/D_{max}),$$

where N = the number of things to be placed in ordered categories, and D_{max} = the maximum value that D can assume. Our initial sim-

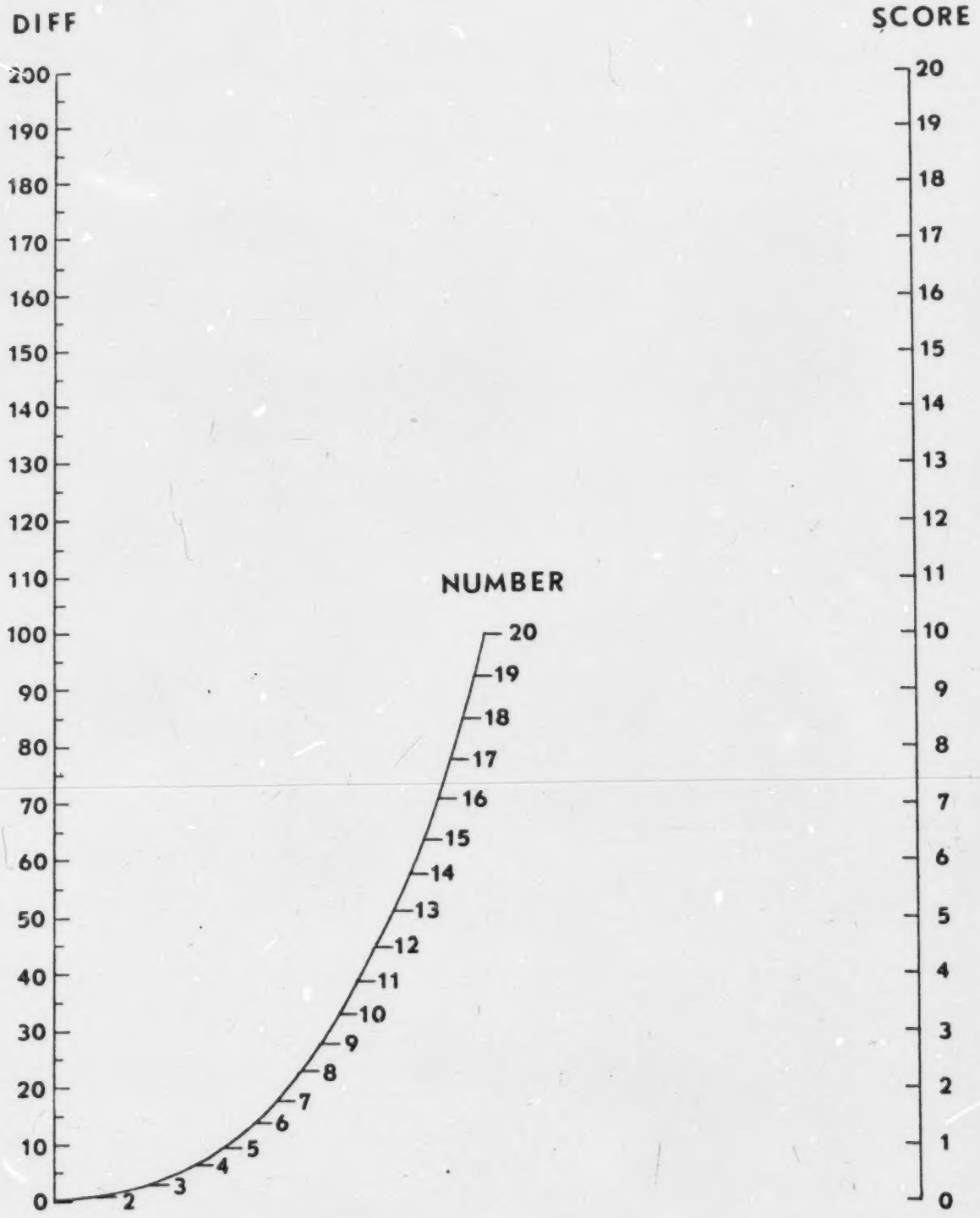

Fig. 1. A nomograph for scoring rearrangement items with one option specified per category. DIFF is the sum of the absolute values of the differences between the correct ordering and the examinee's ordering of the NUMBER of options. SCORE, the examinee's final score, is the SCORE-scale intersection of the line extended through the appropriate points on the DIFF and NUMBER scales.

plifying approach in using this formula was to prepare another nomograph which could be entered with the values of N, D, and D_{max}. As illustrated in Figure 2, a line extended from point "d" on the D scale through the appropriate value of D_{max} will intersect the vertical line Q at point "q". From point "q" a line is then extended through the appropriate value of N on the N scale; this line intersects the S scale at the correct value of S. The problem with scoring ordered multicategory items in this way is the fact that D_{max} must first be computed. Although a manual procedure for doing this was devised, a computer program was eventually written to determine final scores on all types of multicategory items. This program is given in Table 1.

In using the program, a control card appears first in the data deck. This card contains, in three successive four-column fields, the number of categories, the total number of items (options), and the total number of examinees. Following the control card is a card(s) containing, in successive two-column fields for each options, the scoring key for the item. For example, on a three-category item having five options, if the first option belongs in category 2, the second option in category 3, and the last three options in category 1, then on this card a 2 is punched in column 2, a 3 in column 4, and 1's in columns 6, 8, and 10. The card for statement 2 of the program, which must also be prepared, supplies the appropriate format for the scoring key. Thus, for the above example statement 2 of the program would be 2 FORMAT(5I2). Finally, a separate data card for each examinee, containing the numbers of the categories in which he placed each successive option (or item), is punched according to the same format as the

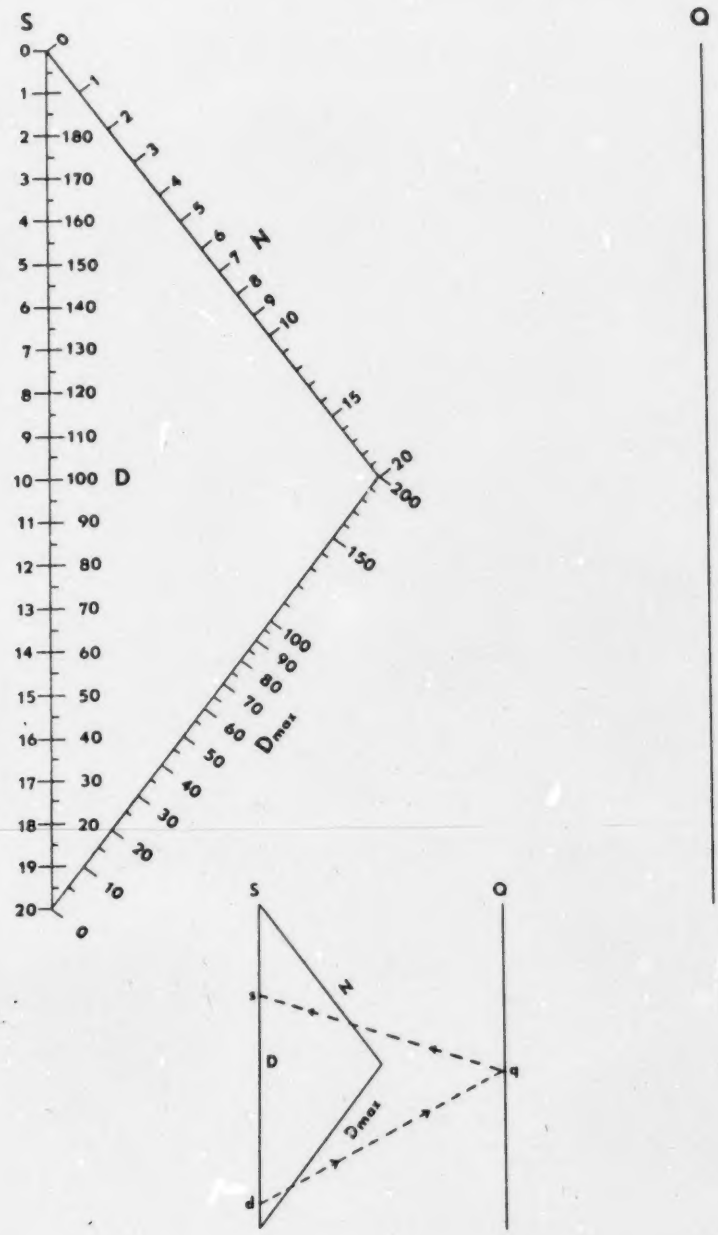

Fig. 2. A nomograph for scoring any type of rearrangement item. D is the sum of the absolute values of the differences between the correct ordering and the examinee's ordering of the N options, N is the total number of options, Dmax is the maximum possible value of D, and S is the final score. To find S, extend a line from points d (the value of D for this examinee) through the appropriate value on the Dmax axis to the Q axis. A line from this point (q) is then extended through the appropriate value on the N axis to the S axis. Point s on the S axis is the score for this examinee.

```
C                              TABLE 1
C          PROGRAM FOR SCORING ORDERED MULTICATEGORY ITEMS
      DIMENSION IY(100),IX(100),M(20)
C     READ NUMBER OF CATEGORIES (NCATS), TOTAL NUMBER OF ITEMS (NITMS),
C     AND NUMBER OF EXAMINEES (NEXMS).
      READ(2,1)NCATS,NITMS,NEXMS
    1 FORMAT(3I4)
C     READ SCORING KEY FOR ITEMS AND COUNT NUMBER OF ITEMS IN EACH CATEGORY.
      READ(2,2)(IY(I),I=1,NITMS)
    2 FORMAT(*****)
      DO 3 K=1,NCATS
    3 M(K)=0
      DO 6 I=1,NITMS
      DO 5 K=1,NCATS
      IF(IY(I)-K)5,4,5
    4 M(K)=M(K)+1
      GO TO 6
    5 CONTINUE
    6 CONTINUE
C     COMPUTE MAXIMUM SUM OF ABSOLUTE VALUES OF DIFFERENCES AND MAXIMUM
C     SUM OF SQUARED DIFFERENCES BETWEEN CORRECT RESPONSE PATTERN AND ANY
C     RESPONSE PATTERN.
      MABS=0
      MSQS=0
      N1=NCATS-1
      DO 12 I=1,N1
      ICAT=NCATS-I
      DO 11 L=1,ICAT
      J=NCATS-L+1
      IF(M(I))12,12,7
    7 IF(M(J))11,11,8
    8 IF(M(I)-M(J))9,9,10
    9 MABS=MABS+2*M(I)*(J-I)
      MSQS=MSQS+2*M(I)*(J-I)**2
      M(J)=M(J)-M(I)
      M(I)=0
      GO TO 12
   10 MABS=MABS+2*M(J)*(J-I)
      MSQS=MSQS+2*M(J)*(J-I)**2
      M(I)=M(I)-M(J)
      M(J)=0
   11 CONTINUE
   12 CONTINUE
      WRITE(3,13)MABS,MSQS
   13 FORMAT('1MAXIMUM SUM OF ABSOLUTE VALUES OF DIFFERENCES =',I4/' MAX
     1IMUM SUM OF SQUARED DIFFERENCES =',I4)
C     READ RESPONSES OF EACH EXAMINEE.  THEN COMPUTE SUM OF ABSOLUTE
C     VALUES OF DIFFERENCES AND SUM OF SQUARED DIFFERENCES BETWEEN
C     CORRECT RESPONSE PATTERN AND EXAMINEE'S RESPONSE PATTERN.
      WRITE(3,14)
   14 FORMAT('0EXAMINEE      SUM OF ABS VALUES OF DIFFS      SCORE      SUM
     1 OF SQUARED DIFFS      CORRELATION')
      DO 16 L=1,NEXMS
      NABS=0
      NSQS=0
      READ(2,2)(IX(I),I=1,NITMS)
      DO 15 I=1,NITMS
      NABS=NABS+IABS(IX(I)-IY(I))
   15 NSQS=NSQS+(IX(I)-IY(I))**2
C     COMPUTE EACH EXAMINEE'S SCORE AND CORRELATION BETWEEN EXAMINEE'S
C     RESPONSE PATTERN AND CORRECT RESPONSE PATTERN.
      SCORE=NITMS*(1.-1.*NABS/MABS)
      CORR=1.-2.*NSQS/MSQS
   16 WRITE(3,17)L,NABS,SCORE,NSQS,CORR
   17 FORMAT(3X,I3,18X,I4,15X,F7.3,12X,I4,15X,F6.3)
      CALL EXIT
      END
```

BEST COPY AVAILABLE

scoring key card.

The output of the program gives the maximum sum of the absolute values of the differences (D_{max}) and the maximum sum of the squared differences. Also, the output for each examinee includes the sum of the absolute values of the differences between his ordering of the options (items) and the correct order (D), the corresponding value of S computed from formula 2, the sum of squared differences, and the correlation between the examinee's ordering of the options and the correct order. The correlation coefficient is a generalized rank-order coefficient, but unlike Spearman's rho, which is a special case of this coefficient, it cannot be derived directly from the Pearson product-moment formula. The range of the generalized rank-order coefficient computed on ordered multicategory data is usually less than that.

The means and standard deviations of scores computed by applying formula 2 to 5-category data on 10 samples of 40 examinees each were determined. These scores were also correlated with scores on the same items computed by a simple count of the number of correspondences between the examinee's ordering of the options and the correct ordering. Both the means and standard deviations of formula 2 scores were significantly higher than those obtained from a correspondences count. The correlations between the two methods of scoring ranged from .583 to .794 for the 5-category data used.

Scoring Dichotomous Items

Since a differential weighting system seemed to make sense on ordered multicategory items, and it did affect the psychometric characteristics of the scores, we next turned out attention to a

reduced ordered multicategory item--the true-false dichotomous item. As proposed by Ebel (1965), one interesting revision of the true-false format is to have examinees indicate the degree of confidence that they have in their answers. If the statement is actually true, the examinee is given 2, 1, 0 or −2 points, depending on whether he indicated that the statement is probably true, possibly true, possibly false, or probably false, respectively. On the other hand, if the statement is actually false, the examinee receives −2 points for answering probably true, 0 points for possibly true, 1 point for possibly false, and 2 points for probably false. Finally, regardless of the truth or falsity of the statement, the examinee receives .5 points if he marks "I have no idea." Ebel reports gains in reliability from using confidence weighting equivalent to what would be expected by a 75 to 80 per cent increase in the length of the test.

Our inspection of numerous true-false and multiple-choice items revealed that all item options were not keyed an equal number of times on all of these tests. On several true-false tests, for example, "true" was correct on 75% or more of the items. And on four-option multiple-choice items, the "b" and "c" options were keyed significantly more often than the "a" and "d" options. Consequently, we became concerned with the effects on test scores of non-uniformity in the distributions of keyed options. We also wanted to look at the differential effects of option keying and other variables by using a variety of test scoring methods. To do this, we began by limiting our investigations to dichotomous items and the seven scoring formulas listed in Table 2. In these formulas,

Table 2

Seven Scoring Formulas for Dichotomous Items

(1) $S_1 = R$ (conventional "number right" scoring formula)

(2) $S_2 = R - W$ (conventional "correction for guessing" formula)

(3) $S_3 = R - \frac{AW_a}{B} - \frac{BW_b}{A}$ (individual categories corrected for guessing)

(4) $S_4 = \frac{N}{2}\left(\frac{R_a}{A} + \frac{R_b}{B}\right)$ (weight for item correct inversely proportional to number of items in category)

(5) $S_5 = \frac{N}{2}\left(\frac{R_a - W_a}{A} + \frac{R_b - W_b}{B}\right)$ (formula 4 corrected for guessing)

(6) $S_6 = N(R - |A - B|)/(N - |A - B|)$ (absolute value of differences)

(7) $S_7 = \left(\frac{N}{N - |A - B|}\right)\left[R - |A - B| - (W - |A - B|)\left(\frac{A^2 + B^2 - N|A - B|}{2AB - N|A - B|}\right)\right]$

(formula 6 corrected for guessing)

S = the final score,

R = the total number of items that the examinee gets right,

W = the total number of items that the examinee gets wrong,

A = the number of items on which option "a" is keyed as correct,

B = the number of items on which option "b" is keyed as correct,

N = A + B,

R_a = the number of A items that the examinee gets right,

R_b = the number of B items that the examinee gets right,

W_a = the number of A options that the examinee gets wrong, and

W_b = the number of B options that the examinee gets wrong.

Formula 1 is the conventional "number right" scoring formula, and formula 2 is the conventional correction for guessing formula for dichotomous items. Formula 3 is formula 1 with the a-option-correct and b-option-correct items individually corrected for guessing. In formula 4, the weight for answering an item correctly is inversely proportional to the number of items in the category (a-option-correct or b-option-correct) to which the item belongs; formula 5 is formula 4 corrected for guessing. Formula 6 is an "absolute difference formula" (see Aiken, 1974) in which the score is weighted by the absolute value of the difference between the number of items in the a-option-correct category and the number of items in the b-option-correct category. Formula 7 is formula 6 corrected for guessing.

When A = B, formulas 4 and 6 reduce to formula 1, and formulas 3, 5, and 7 reduce to formula 2. Consequently, formulas 3-7 are of interest only on double-choice tests where $A \neq B$. Since formulas 1 and 2 are obviously easier to work with than formulas 3-7, we decided to study the empirical characteristics of formulas 3-7 and

to determine whether they possess any advantages at all over the simplicity and convenience of the two conventional formulas.

An Experiment on Seven Methods of Scoring Dichotomous Items

As depicted in Figure 3, the experimental design involved four independent and seven dependent variables. The dependent variables were the seven scoring procedures described in Table 2. The first independent variable (P) was the proportion of the total number of items on which option "a" was keyed as correct. A second independent variable (T) was whether or not examinees were told "P". A third independent variable (G) was whether examinees were instructed to guess when in doubt about the correct answer, and a fourth independent variable (K) was the examinee's knowledge of the subject-matter of the test. To assess the effects of these four independent variables on scores computed from the seven scoring formulas, a comprehensive final examination in general psychology was constructed; the test consisted of 100 double-choice items. These 100 items were the basis of three different item-keying arrangements. In the first arrangement, "a" was the keyed option on 33 items and "b" on the remaining 67 items. In the second arrangement, "a" was the keyed option on 50 items and "b" on the remaining 50 items. In the third arrangement, "a" was the keyed option on 67 items and "b: on 33 items. Furthermore, four different types of instructions were prepared as cover sheets for the test. On half of the instructional sheets examinees were told how many "a" options were correct and how many "b" options were correct; on the other half they were not told. On half of the instructional sheets examinees were told to guess when in doubt; on the other half they were given no instructions regarding guessing.

The effects of the fourth independent variable--examinee's knowledge of the subject-matter of the test--were determined by administering the tests to one group of 84 students (the "high-knowledge" group) during the last week of a general psychology course, and a separate group of 96 students (the "low-knowledge" group) during the first week of a general psychology course. The overall design of the experiment is illustrated by the 3 x 2 x 2 diagram in Figure 3.

After preliminary instructions, in which plausible reasons for cooperating and doing as well as possible on the test were given, the 12 different types of test booklets and the standard answer sheets were distributed in a prearranged manner to the students. When the students had finished the test, they inserted their answer sheets in the test booklets and gave them to the examiner. The "high-knowledge" group was tested in December, 1973 and the "low-knowledge" group in January, 1974. There were eight examinees in each of the 12 low-knowledge groups and seven examinees in each of the 12 high-knowledge groups.

Results of the Experiment

Item responses for all examinees were punched on IBM cards and then scored and analyzed by correlational, analysis of variance, and factor analysis procedures. The correlation matrices for the seven formulas applied to the scores of the low-knowledge and high-knowledge groups are given in Table 3. The F ratios obtained from the seven analyses of variance, one for each of the seven formulas, for the low-knowledge group are given in the first half of Table 4. The corresponding F ratios for the high-knowledge group are given in the second half of Table 4. The group means

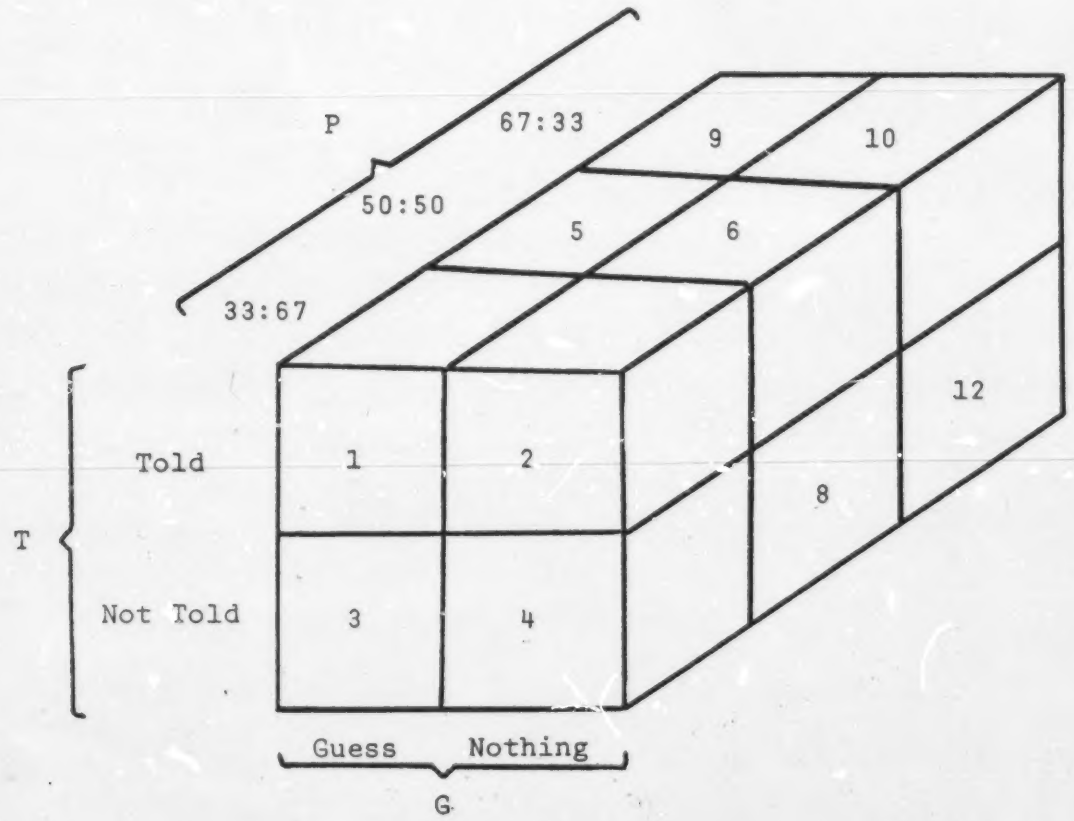

Fig. 3. Design of the experiment on seven methods of scoring dichotomous items. P is the ratio of the number of items on which "a" is the correct option (A) to the number of items on which"b" is the correct option. T is whether or not examinees were told P. G is whether examinees were told to guess when in doubt or were given no instructions concerning guessing. The numbers in the boxes are the labels for the twelve experimental groups.

Table 3

Matrices of Correlations Among Scores from Seven Formulas

Low Knowledge Groups (N = 96)

	Formula						
	1	2	3	4	5	6	7
1	1.00						
2	.67	1.00					
3	.40	.92	1.00				
4	.98	.68	.40	1.00			
5	.56	.96	.90	.62	1.00		
6	.80	.94	.83	.83	.90	1.00	
7	.64	.99	.93	.66	.96	.93	1.00

High Knowledge Groups (N = 84)

	Formula						
	1	2	3	4	5	6	7
1	1.00						
2	.99	1.00					
3	.88	.88	1.00				
4	.98	.98	.84	1.00			
5	.98	.98	.84	.99	1.00		
6	.82	.82	.92	.81	.81	1.00	
7	.73	.73	.50	.71	.70	.41	1.00

Table 4

F Ratios for the Seven Scoring Formulas

Variable	Formula 1	Formula 2	Formula 3	Formula 4	Formula 5	Formula 6	Formula 7
			Low-Knowledge Group				
P	.83	.64	14.39**	.70	.51	44.50**	2.78
T	3.74*	4.50*	8.49**	1.56	2.02	5.51*	6.53*
G	6.81*	7.94**	8.15**	7.27**	8.38**	8.95**	10.47**
P X T	.83	.90	1.84	.39	.44	1.37	1.43
P X G	.59	.63	.79	.67	.71	1.19	1.44
T X G	5.16*	4.12*	4.48*	4.81*	3.84	5.54*	4.81*
P X T X G	.19	.21	.43	.15	.16	.47	.51
			High-Knowledge Group				
P	.06	.07	8.00**	.05	.06	21.39**	7.69**
T	2.01	2.01	4.48*	.56	.56	2.80	3.05
G	.67	.75	.94	.75	.83	1.21	1.47
P X T	2.09	2.12	2.99	1.38	1.41	2.34	2.40
P X G	1.53	1.53	1.51	1.50	1.50	1.50	1.44
T X G	.59	.61	.15	.90	.92	.16	.16
P X T X G	1.47	1.65	1.49	1.24	1.39	1.00	.94

*Significant at the .05 level.
**Significant at the .01 level.

are given in Table 5.

As indicated by the number of significant F ratios in the first half of Table 4 compared to the second half, the effects of variables T and G were apparently greater in the low-knowledge than in the high-knowledge group. Whenever variable T was significant, the group that was told A and B, and to guess when in doubt, typically scored higher than the group not told these numbers but told to guess when in doubt. It should also be noted that, for a given combination of P and T, the group instructed to guess when in doubt about the correct answer to an item usually scored lower than the group given no instructions whatsoever concerning guessing. Finally, in both the low-knowledge and high-knowledge groups, the ratio of A to B had a greater effect on scores computed by formulas 3, 6, and 7 than on those computed by formulas 1, 2, 4, and 5.

Conclusions

In general, it appears that different instructions and different values of A and B have greater effects on scores computed from formulas 3, 6, and 7, and especially with examinees having low knowledge of the subject-matter of the test. It is clear from the results that whenever formulas 3, 6, or 7 are used, examinees should certainly be informed of the ratio of "a" options keyed to "b" options keyed. On the other hand, the experimental variables had less effect on scores obtained from formulas 1, 2, 4, and 5, especially in the high-knowledge group. They did, however, have significant effects on these scores in the low-knowledge group. If one views less sensitivity to the effects of instructions as an advantage of a scoring formula, then formulas 1 and 4 and

Table 5

Means of Twelve Experimental Groups in Low-Knowledge and High-Knowledge Sections

Experimental Condition			Formula						
R	T	G	1	2	3	4	5	6	7
Low-Knowledge Group (N = 104)									
1:2	Told	Guess	63.50	27.00	11.49	61.70	23.41	44.69	39.90
1:2	Told	---	64.75	30.50	14.19	63.53	28.07	46.57	46.70
1:2	Not Told	Guess	54.25	8.50	-16.34	54.70	9.42	30.68	8.15
1:2	Not Told	---	65.38	30.75	11.22	66.08	32.17	47.53	46.34
1:1	Told	Guess	61.50	23.00	23.00	61.50	22.99	61.50	23.00
1:1	Told	---	60.88	22.50	22.44	60.89	22.52	60.65	22.56
1:1	Not Told	Guess	58.75	17.50	17.50	58.75	17.49	58.75	17.50
1:1	Not Told	---	63.25	26.50	26.50	63.25	26.49	63.25	26.50
2:1	Told	Guess	61.38	22.75	4.29	60.50	21.10	41.47	32.60
2:1	Told	---	62.38	27.25	10.21	61.78	26.07	44.38	40.28
2:1	Not Told	Guess	53.12	7.12	-17.20	53.12	7.13	29.24	5.92
2:1	Not Told	---	61.00	22.12	.42	61.60	23.33	40.85	31.82
High-Knowledge Group (N = 85)									
1:2	Told	Guess	69.57	39.14	25.58	68.39	36.79	53.89	60.74
1:2	Told	---	72.57	45.14	32.48	71.73	43.46	58.44	71.04
1:2	Not Told	Guess	66.17	32.33	13.11	66.93	33.86	48.73	49.05
1:2	Not Told	---	72.40	45.20	30.29	72.87	46.14	58.56	70.78
1:1	Told	Guess	68.43	36.86	36.86	68.43	36.85	68.43	36.86
1:1	Told	---	71.14	43.00	43.00	71.14	42.99	71.10	43.05
1:1	Not Told	Guess	75.14	50.71	50.70	75.15	50.72	75.08	51.02
1:1	Not Told	---	67.43	35.00	34.99	67.43	35.00	67.38	35.05
2:1	Told	Guess	72.14	44.57	32.71	70.95	42.19	58.05	69.78
2:1	Told	---	74.43	49.00	35.80	74.23	48.61	61.14	77.76
2:1	Not Told	Guess	65.57	31.14	10.83	66.72	33.45	47.83	47.01
2:1	Not Told	---	67.29	34.86	16.22	68.04	36.38	50.48	53.46

formulas 2 and 5 are quite comparable and better than formulas 3, 6, and 7. It is also interesting that a principal axis factor analysis of these seven formulas yielded two negatively correlated ($r = -.668$) factors for the low-knowledge data but only one factor for the high-knowledge data (see Table 6). Factor 1 for the low-knowledge data had its highest loadings on formulas 2, 3, 5, and 7, whereas factor 2 had its highest loadings on formulas 1 and 4. It can be observed in Table 2 that formulas 2, 3, 5 and 7 include a correction for guessing, while formulas 1 and 4 do not. The obvious implication of this fact, coupled with the significance of G for the low-knowledge but not the high-knowledge groups in Table 4, is that guessing, and consequently "correction for guessing" have a greater influence on test scores when examinees are less informed on the subject matter of the test.[1]

We are currently conducting a further investigation of these seven formulas, including an assessment of their acceptability to examiners and examinees and their usefulness on other types of tests. Preliminary results indicate that examinees prefer the simpler formulas--the ones that they understand. It is obviously important that the people involved understand the basis on which a test is being scored, and in this respect formulas 1 and 2 are probably superior to the remaining five.

As illustrated in Table 7, formulas 1 through 5 of Table 1 can be generalized to multiple-choice items having k options each. In these formulas, i = the number of the option ($i = 1, k$), a_i = the total number of items having i as the keyed option, r_i = the number of correctly-answered items having i as the correct option,

Table 6

Factor Loadings On Seven Formulas for Low Knowledge and High Knowledge Groups

Low Knowledge Group (N = 96)

Scoring Formula	Factor Loadings	
	Factor 1	Factor 2
1	-.053	.790
2	.710	.046
3	.886	-.284
4	.053	.658
5	.756	-.053
6	.510	.284
7	.738	.007

High Knowledge Grcup (N = 84)

Scoring Formula	Factor Loadings
	Factor 1
1	.996
2	.994
3	.902
4	.985
5	.986
6	.860
7	.705

Table 7

Five Scoring Formulas for Multiple Choice Items

(1) $S_1 = \sum_{i=1}^{k} r_i$

(2) $S_2 = \sum_{i=1}^{k} (r_i - w_i)$

(3) $S_3 = \sum_{i=1}^{k} (r_i - \frac{a_i w_i}{N - a_i})$

(4) $S_4 = \frac{N}{k} (\sum_{i=1}^{k} r_i/a_i)$

(5) $S_5 = \frac{N}{k} \sum_{i=1}^{k} (\frac{r_i}{a_i} - \frac{w_i}{N - a_i})$

S_j = final score using the jth formula

k = number of options per item

i = ith option

a_i = total number of items having i as keyed option

r_i = number of correctly-answered items with i as correct option

$N = \sum_{i=1}^{k} a_i$

$w_i = N - r_i$

$N = \sum_{i=1}^{k} a_i$, and $w_i = N - r_i$. We are currently planning comparable investigations with these and other formulas for scoring multiple-choice tests having more than two options per item.

[1]Quereshi (1974) reported evidence that informing college student examinees that there would be a penalty for guessing on a test decreased the number of right answers and omissions, but increased the number of wrong answers.

References

Aiken, L. R. Scoring for partial knowledge on the generalized rearrangement item. Educational and Psychological Measurement, 1970, 30, 87-94.

Aiken, L. R. Procedures for computing scores and correlations on ordered multicategory test items. Educational and Psychological Measurement, 1974, 34, 173-176.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Quereshi, M. Y. Performance on multiple-choice tests and penalty for guessing. Journal of Experimental Education, 1974, 42, 74-77.

Stanley, J. C., & Wang, M. D. Weighting test items and test-item options: An overview of the analytical and empirical literature. Educational and Psychological Measurement, 1970, 30, 21-35.